Gated Fusion-in-Decoder를 이용한 대화 시스템 응답 성능 개선

Gated Fusion-in-Decoder for Improving Response Generation in Dialogue System

전북대학교 JEONBUK NATIONAL UNIVERSITY

정민교·나승훈(전북대학교)·홍범석·최원석·한영섭·전병기(LG 유플러스)

{ jungmk, nash } @ jbnu.ac.kr · {bshong, wseokchoi, yshan042, bkjeon} @ lguplus.co.kr

Ⅰ. 서 론

연구 목적

사용자 발화와 연관된 문서를 검색하고, 검색된 문서가 그라운딩된 응답을 생성하는 문서 그라운딩 대화 시스템(Document grounded Dialogue system) 태스크의 성능 개선을 위해 본 연구를 수행하였다.

문서 그라운딩 대화 시스템은 검색, MRC(Machine Reading Comprehension) 두 단계로 수행된다. 다양한 MRC 방법론 중 FiD(Fusion-in-Decoder) 모델은 검색된 다중 문서를 사용하여 답변 생성 성능이 향상된 것을 보였다. 하지만 주어진 질의와 검색된 다중 문서간의 연관 정도가 각각 다름을 활용하기 어렵다는 한계가 있다. 본 논문에서는 검색된 문서간의 Gating mechanism을 통해 문서간 가중치를 학습하고, FiD 모델과 비교해 응답 생성 성능을 향상하였다.

Ⅱ. 제안 모델

모델 구조

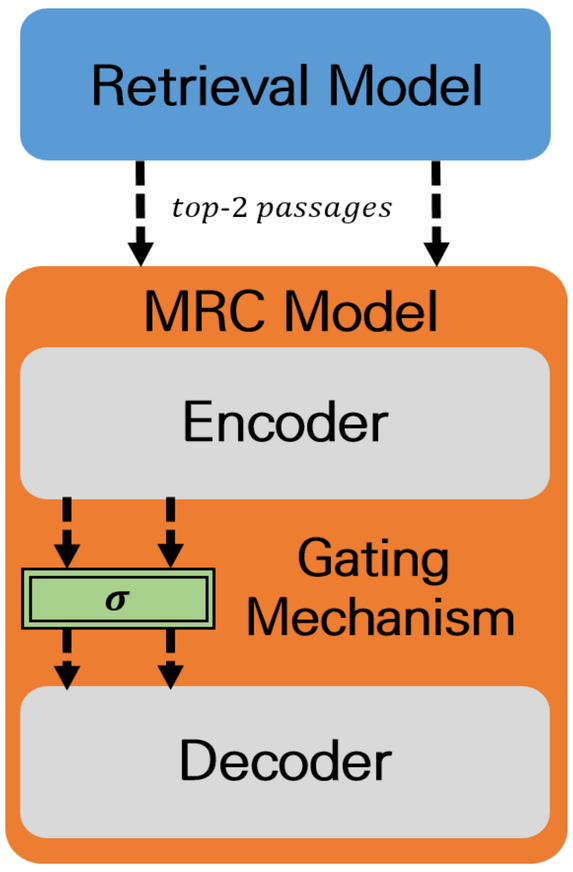

그림 1: Model Architecture

Dens Passage Retrieval + Re-ranking

$$prompt_i = History: d;\ Context: c_i \quad (1)$$

$$input\_ids = [prompt_1, prompt_2] \quad (2)$$

T5-base

$$h_i = Encoder(input\_ids) \quad (3)$$

$$h = gating(h_1, h_2) \quad (4)$$

$$\alpha = \sigma(Linear_1(x_1) + Linear_2(x_2)) \quad (5)$$

$$o = [\alpha \odot x_1; (1-\alpha) \odot x_2] \quad (6)$$

Retrieval Model

문서 그라운딩된 대화 시스템 태스크 수행을 위해 주어진 대화 히스토리와 연관된 passage(문서) 검색을 선행하였다. DPR 모델을 사용해 각 대화 히스토리와 연관된 top-k개의 passage를 검색하였고, 검색 성 능을 향상하기 위해 재순위화 작업을 수행하였다.

MRC Model

Retrieval Model을 통해 검색된 상위 2개의 passage는 수식 (1)과 같이 각 'History:', 'Context:' 프롬프트 뒤에 히스토리와 검색된 passage 내용이 작성된 후 모델 입력을 위한 token id로 변환되어 $prompt_i$를 구성한다. 이 후, 모델 입력을 위해 $input_ids$로 결합한다.

$input_ids$는 MRC 모델의 인코더를 통과해 수식 (3)과 같이 인코더 아웃풋이 passage 단위로 분리되어 h_i로 출력된다. 이 후 수식 (4)와 같이 Gating mechanism을 통해 각각 가중치가 곱해지고, 결합되어 h로 출력된다.

Gating mechanism은 수식 (5), (6)과 같이 두 인자 x_1, x_2를 입력 받고 각 선형층과 시그모이드 함수를 통해 가중치 α를 출력하는 과정으로 이루어진다. Gating mechanism을 통해 출력된 h는 디코더를 통과해 최종 아웃풋으로 출력된다.

Ⅲ. 실험

데이터셋

태스크 수행을 위해 MultiDoc2Dial 데이터셋을 사용하였다. MultiDoc2Dial 데이터셋은 Seen-data와 Unseen-data로 구성되어 있어 본 실험에서는 Seen-data로 학습과 테스트를 수행하였고, Unseen-data는 테스트만 수행하였다. 이 때, Seen-data, Unseen-data는 각각 29,746개, 121개 질의와 3,820개, 100개의 passage로 구성되어 있다.

Baseline model

본 연구에서는 제안 모델의 성능 향상을 확인하기 위해 동일한 Retrieval Model의 검색 결과를 사용하였고, MRC Model은 T5-base를 사용하였다.

Baseline Model은 FiD 모델이고, 검색된 top-2개의 passage를 입력으로 주어진 질의에 따른 답변을 생성하였다.

실험 결과

본 연구에서는 생성된 답변의 성능 지표로 token-level F1 Score와 Rouge-L, SacreBleu matric을 사용하고, 세 지표의 합산 점수를 기준으로 모델 별 성능을 비교하였다.

표 1: 문서 그라운딩 된 대화 시스템 태스크 실험 결과

Dataset type	Model	F1 Score	Rouge-L Score	SacreBleu Score	Total Score
Seen-data (gold / retrieved)	Baseline (ours)	61.29 / 44.78	57.29 / 41.32	39.09 / 25.86	157.67 / 111.96
	Proposed	**61.51 / 44.96**	**57.39 / 41.45**	**39.43 / 26.14**	**158.33 / 112.55**
Unseen-data	Baseline (ours)	29.78	27.56	13.98	71.32
	Proposed	**30.31**	**27.75**	13.74	**71.80**

Seen-data의 결과의 좌측에 제시된 성능은 gold passage를 top-1 passage로 사용한 응답 생성 성능이고, 우측에 제시된 성능은 검색된 passage를 사용한 응답 생성 성능이다.

Seen-data를 사용한 실험에서는 gold passage(gold)를 사용했을 때 제안 모델의 성능이 Baseline Model과 비교해 약 0.66점의 Total Score가 향상되었고, 검색된 passage(retrieved)를 사용했을 때 제안 모델의 성능이 Baseline Model과 비교해 약 0.59wja의 Total Score가 향상된 것을 확인하였다.

Unseen-data를 사용한 실험에서는 제안 모델의 성능이 Baseline Model과 비교해 약 0.48점의 Total Score가 향상된 것을 확인하였다.

Ⅳ. 결 론 및 향후 연구

결 론

문서 그라운딩된 대화 시스템 태스크의 응답 생성 성능 향상을 위해 FiD 모델에 Gating mechanism을 추가 하였다. Seen-data, Unseen-data 모두에서 변화량이 작지만 성능이 향상된 것을 확인 하였다. 이는 주어진 대화 히스토리에 적절한 답변 생성을 위해 Gating mechanism이 입력된 두 passage간의 가중치를 조절해 성능 향상에 영향을 끼쳤다는 것을 의미한다.

향후 연구

본 연구에서는 검색된 passage들 중 상위 2개의 passage만을 사용한 한계를 보였다. FiD 모델의 실험 결과에 따라 추가적인 passage를 사용한다면 성능이 향상될 수 있을 것으로 기대하고, 검색된 여러 passage들 간의 중요도를 학습할 수 있는 연구를 향후에 수행하고자 한다.